Estimation of gene regulation using expression profiles by gene disruption and comprehensive sequence analysis on gene regulatory regions

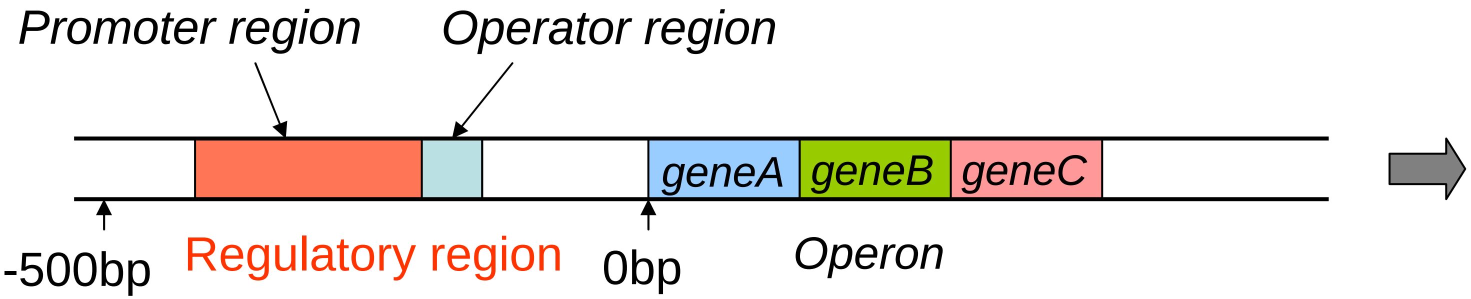
Kunihiko Hiraishi, JAIST and

Hirofumi Doi, Celestar Lexico-Sciences, Inc.

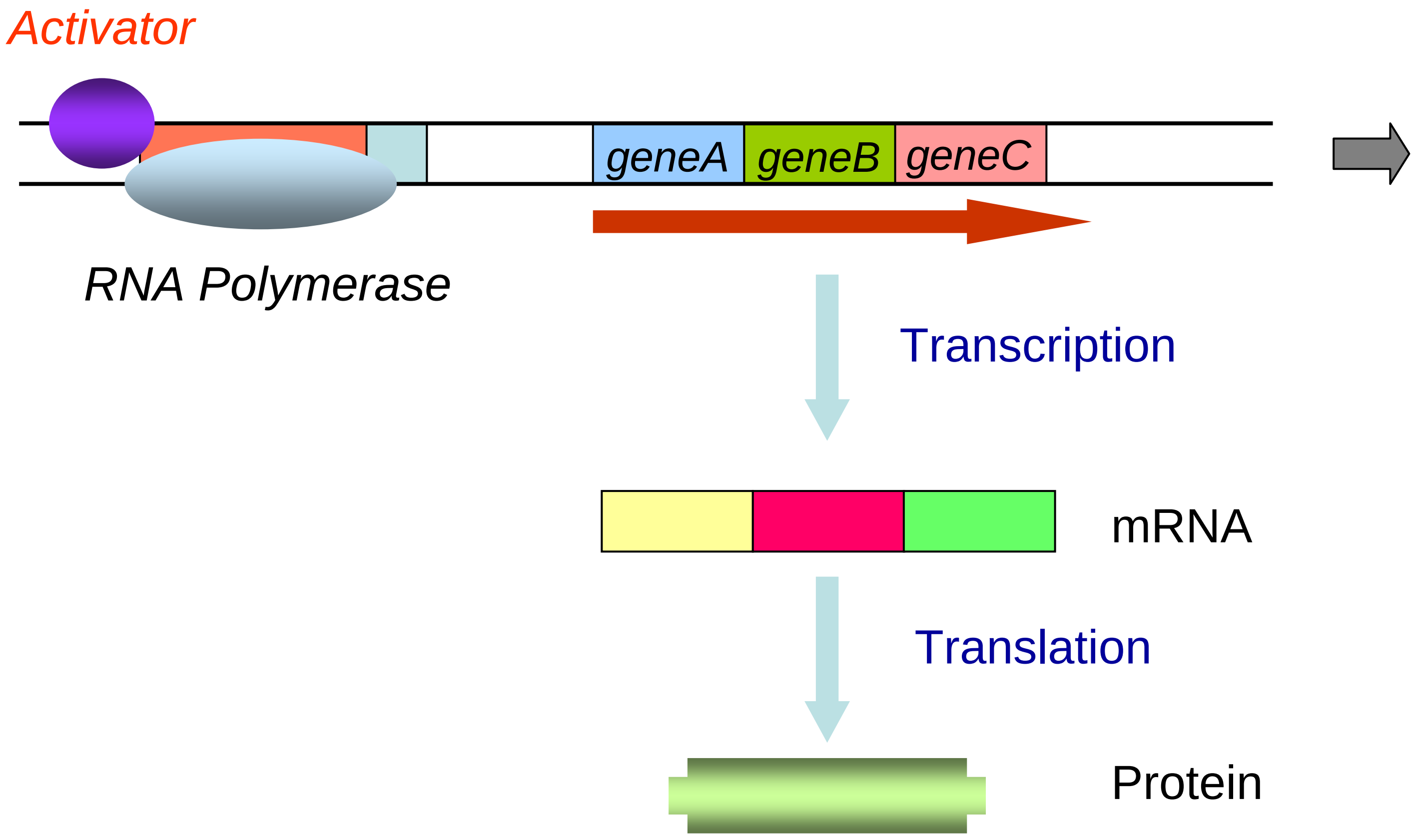
Introduction

- We propose a novel approach to the estimation of gene regulation.
- *The method is a pile of heuristics*.

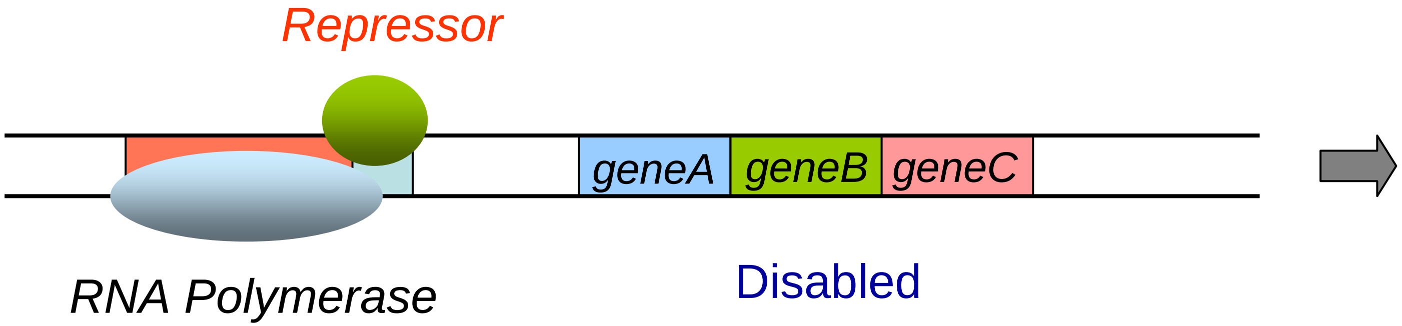
Gene Expression Mechanism



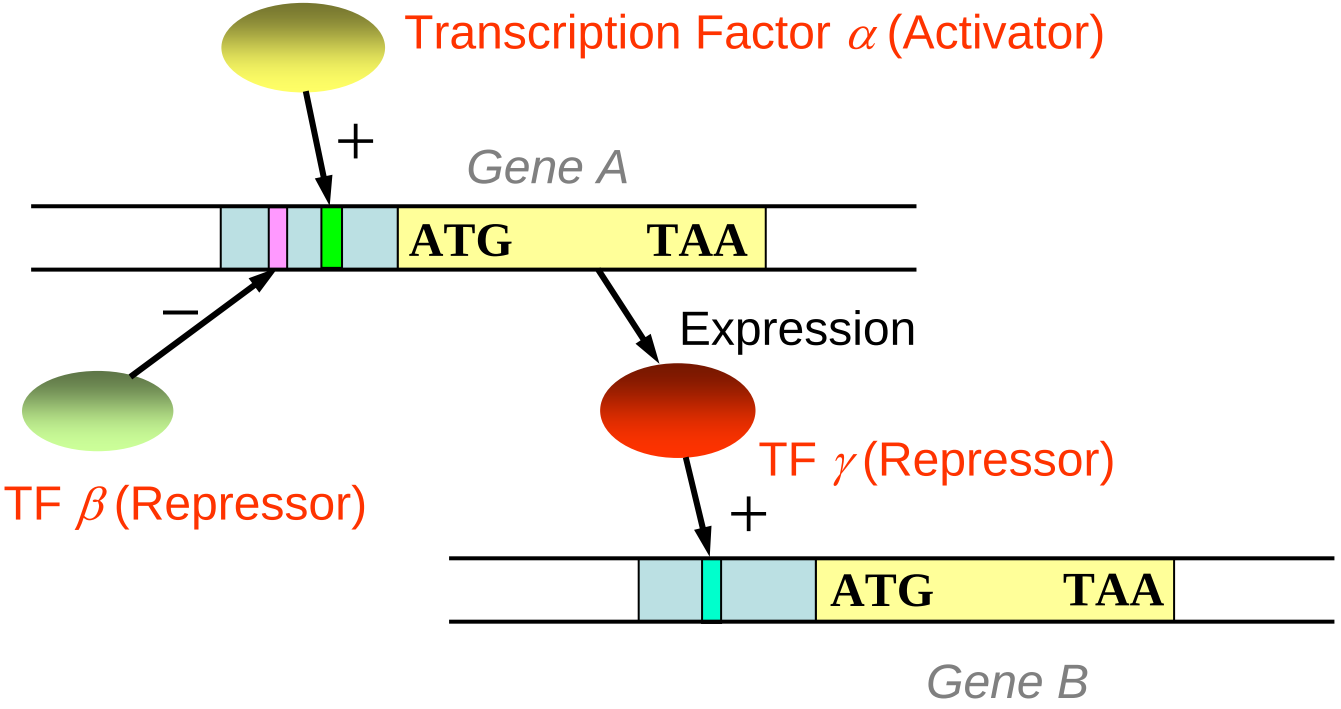
Gene Expression Mechanism



Gene Expression Mechanism



Gene Regulatory Network

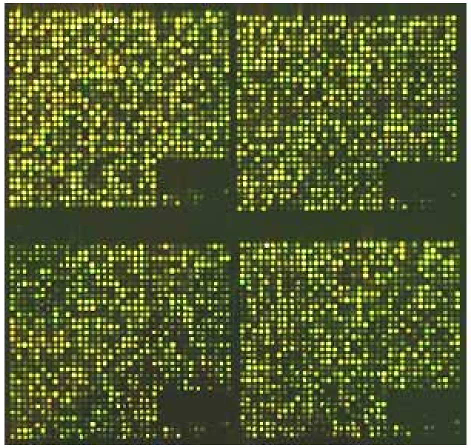


Problem Statement

- Given
  - a set of gene expression data obtained by *disruption of genes*,
  - the complete genome sequence, including the absolute position of each gene,
  - a target gene g_0
- Find
  - a set of genes G_F including g_0 coregulated by *a transcription factor F*, where F is synthesized by one of disrupted genes,
  - *the binding site* of each factor F in the regulatory region of each gene in G_F.

DNA Microarrays

DNA microarrays are used for measuring the expression levels of large numbers of genes simultaneously.



The expression data of gene g_t is a vector *exp*$_t$ such that

$$exp_t[k] = log(M_k / W),$$

where M_k is the expression level of the gene f_k-disruption mutant, and W is the expression level of the wild type.

Difficulties

- Expression profiles by DNA microarray are noisy and errornous.
- Any fluctuations in the expression levels of regulated genes may not be detectable against background fluctuation levels.
- How to identify direct or indirect regulations.

Idea

- If gene g_0 is regulated by factor F synthesized by gene f, then the following holds:
  - The expression profiles between gene g_0 and other co-regulated genes are *correlated*.
  - The expression level of gene g_0 in gene f-disruption mutant *changes significantly*.
  - Gene g_0 and other co-regulated genes *have similar sequence* patterns in their regulatory regions.
- We find a set of genes having all of the three properties (*combination of three independent facts*).
- Using statistical analysis on the frequency of oligonucleotides in regulatory regions, we identify over-represented sequences which may not contribute to the binding of transcription factors, and exclude them from the evaluation of the sequence similarity.

Outline of the Procedure

Step 1. Find a set of genes G whose expression patterns are *correlated* with that of g_0.

Step 2. Compute *window similarity* $w\text{-}sim(w[g_0, i], w[g_t, j])$ for every gene $g_t \in G$ and every positions i, j.

Step 3. Compute subregions R_r on the regulatory region of g_0 such that (i) $max_t\ w\text{-}sim^*(w[g_0, i], g_t)$ is significantly high for almost all $i \in R_r$, and (ii) R_r contains *peak positions* frequently, where $w\text{-}sim^*(w[g, i], g') = max_j\ w\text{-}sim(w[g, i], w[g', j])$.

Step 4. Find a set of transcription factors T_F *dominant* for gene g_0.

Step 5. For each factor $F_k \in T_F$ and each subregion R_r, compute a set of pairs of windows $TFBS(R_r, F_k) = \{ (w[g_0, i], w[g_t, j]) \}$ such that (i) F_k is dominant for gene g_t, (ii) $w\text{-}sim(w[g_0, i], w[g_t, j]) = w\text{-}sim^*(w[g_0, i], g_t)$, and (iii) i is a peak position for g_t.

Outline of the Procedure

Step 1. Find a set of genes G whose expression patterns are correlated with that of g_0.

Step 2. Compute *window similarity* $w\text{-}sim(w[g_0, i], w[g_t, j])$ for every gene $g_t \in G$ and every positions i, j.

Step 3. Compute subregions R_r on the regulatory region of g_0 such that (i) $max_t\ w\text{-}sim^*(w[g_0, i], g_t)$ is significantly high for almost all $i \in R_r$, and (ii) R_r contains *peak positions* frequently, where $w\text{-}sim^*(w[g, i], g') = max_j\ w\text{-}sim(w[g, i], w[g', j])$.

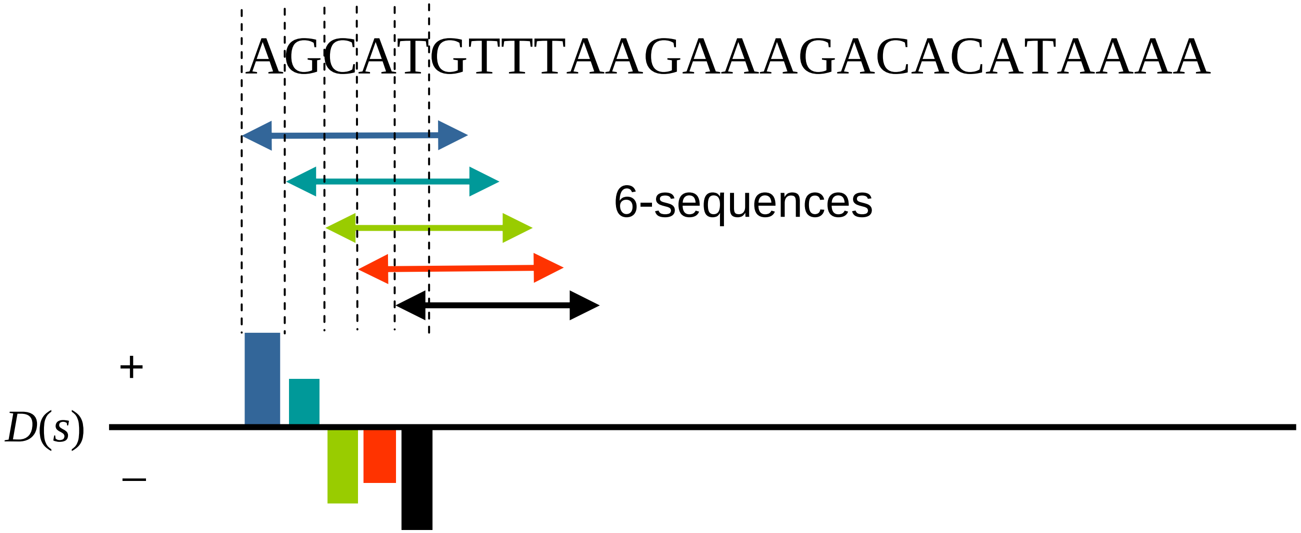
Step 4. Find a set of transcription factors T_F *dominant* for gene g_0.

Step 5. For each factor $F_k \in T_F$ and each subregion R_r, compute a set of pairs of windows $TFBS(R_r, F_k) = \{\ (w[g_0, i], w[g_t, j])\ \}$ such that (i) F_k is dominant for gene g_t, (ii) $w\text{-}sim(w[g_0, i], w[g_t, j]) = w\text{-}sim^*(w[g_0, i], g_t)$, and (iii) i is a peak position for g_t.

Statistical Analysis of Oligonucleotides

- Known binding sequences are often short sequences (around 6bp) or repetition of them with some gap. We call each sequence of length 6 *a 6-sequence*.
- Assumption: *the binding sequences may have some singularity comparing with other sequences*.
- Let $D(s) = log(O_s / E_s)$, where O_s is the actual number of times a 6-sequence s happens in regulatory regions, and E_s is the expected number of times.
- We assume that these 6-sequences with high $D(s)$ do not contribute to binding sequences. This is validated by known binding sequences.

Statistical Analysis of Oligonucleotides

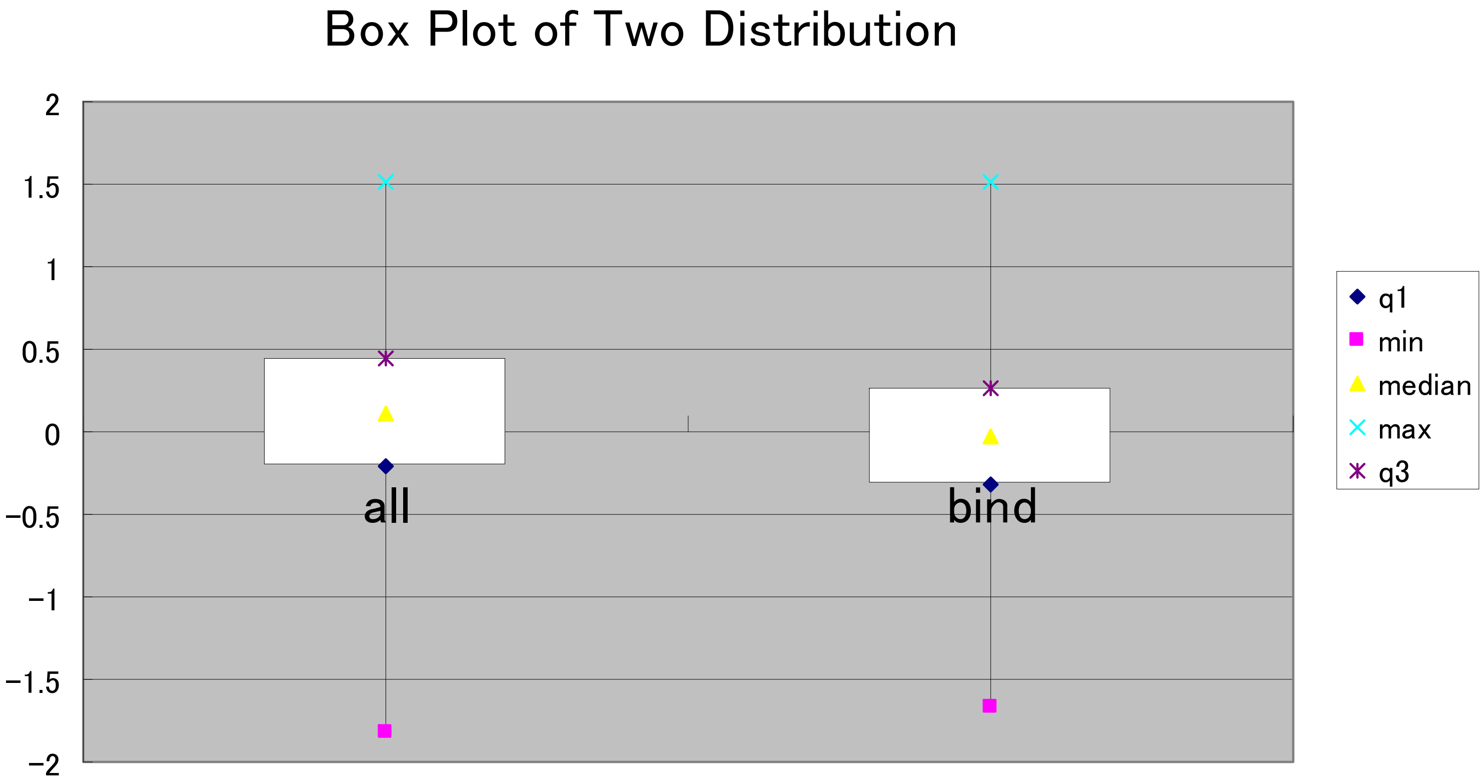


Statistical Analysis of Oligonucleotides

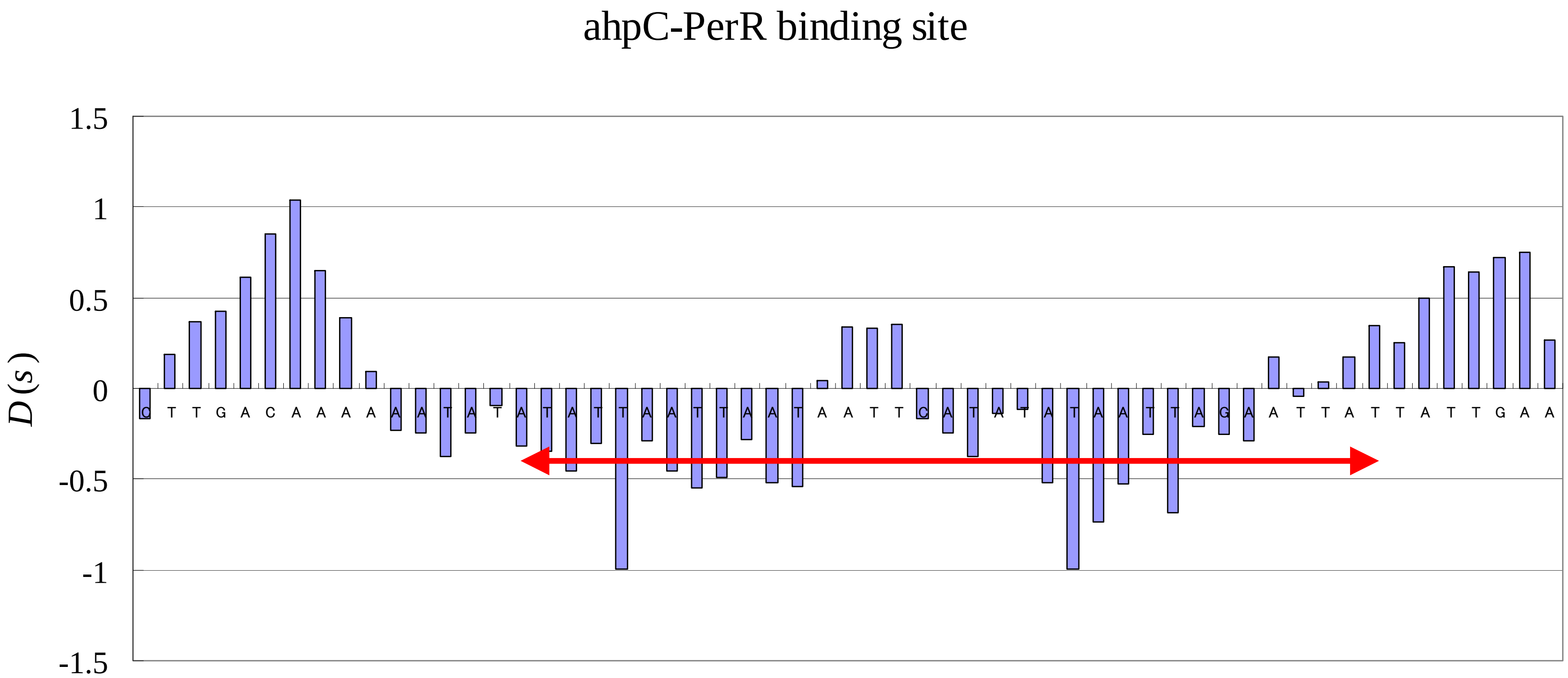
6-equence	O_s	$D(s)$
TTTTTT	3400	1.53397
CTTTTT	2039	1.488083
TTTTTC	1874	1.403698
GGCGGC	345	1.348319
CGGCGG	336	1.321886
GCCGGC	278	1.304955
GCCGCT	371	1.300666
AGGAGG	1026	1.28749
CAGCTG	623	1.243657
CTGCTG	558	1.243394

10 highest 6-sequences

Statistical Analysis of Oligonucleotides

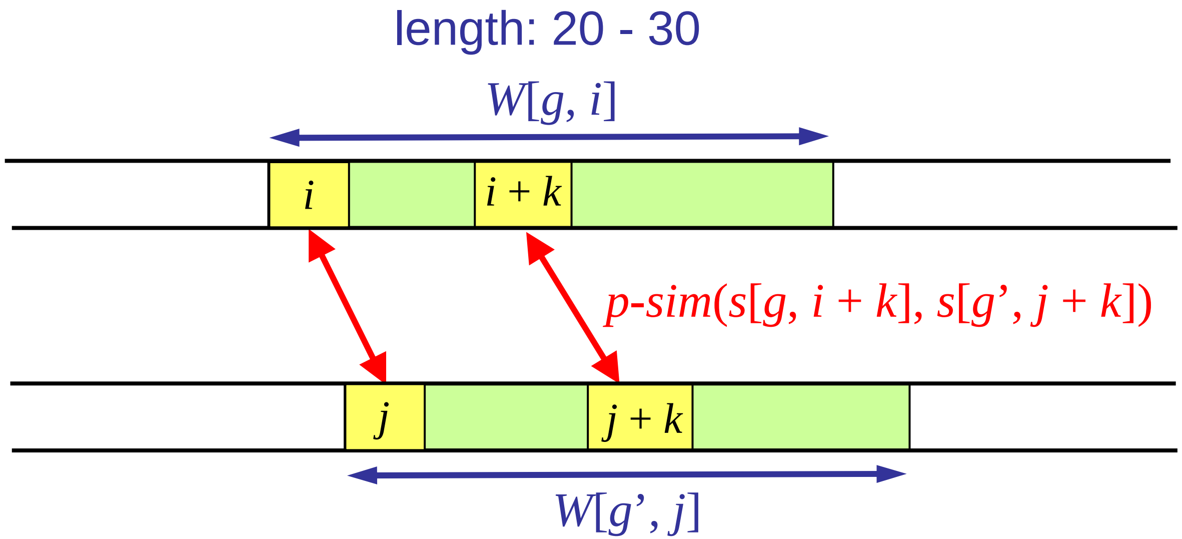


Statistical Analysis of Oligonucleotides



Window Similarity

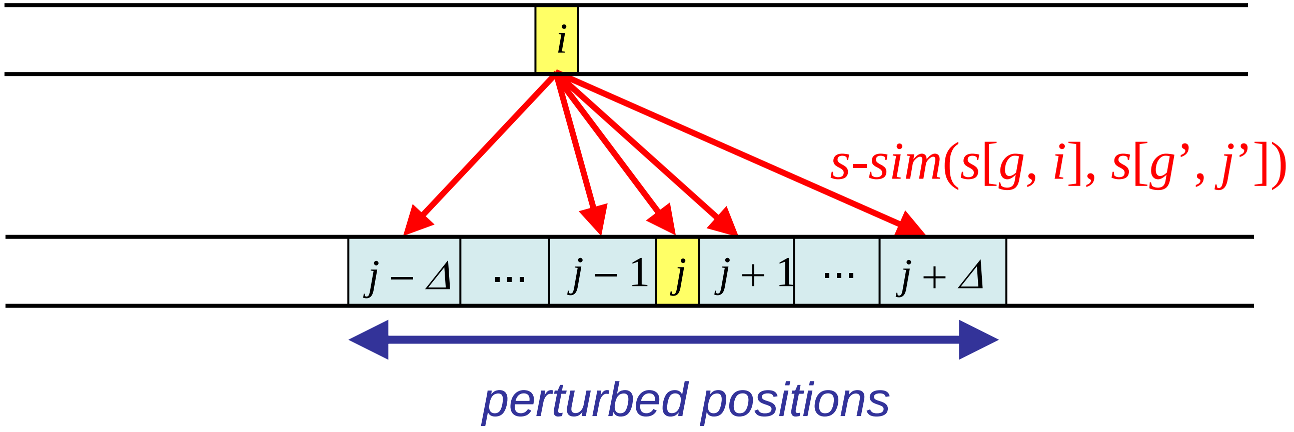
- *Window similarity* = the sum of similarities for pairs of positions (i, j).



$$w\text{-}sim(W[g, i], W[g', j]) := \Sigma_k\, p\text{-}sim(s[g, i + k], s[g', j + k])$$

Window Similarity

- Similarity for a pair of positions (i, j) = maximum *sequence similarity* in all *perturbed positions*.
- *Sequence similarity* = a strictly increasing function of the number of *matched positions*, e.g., $s\text{-}sim(s_1, s_2) := ((1/4)^k(3/4)^{6-k})^{-1}$.
- $s\text{-}sim(s_1, s_2) = 0$ if either s_1 or s_2 is with high $D(s)$ value.
- *Matched positions* : s_1 = **ATTCGT**, s_2 = **AATGGT** $\Rightarrow k = 4$.



$$p\text{-}sim(s[g, i], s[g', j]) = max_{j'}\ s\text{-}sim(s[g, i], s[g', j'])$$

Outline of the Procedure

Step 1. Find a set of genes G whose expression patterns are correlated with that of g_0.

Step 2. Compute *window similarity* $w\text{-}sim(w[g_0, i], w[g_t, j])$ for every gene $g_t \in G$ and every positions i, j.

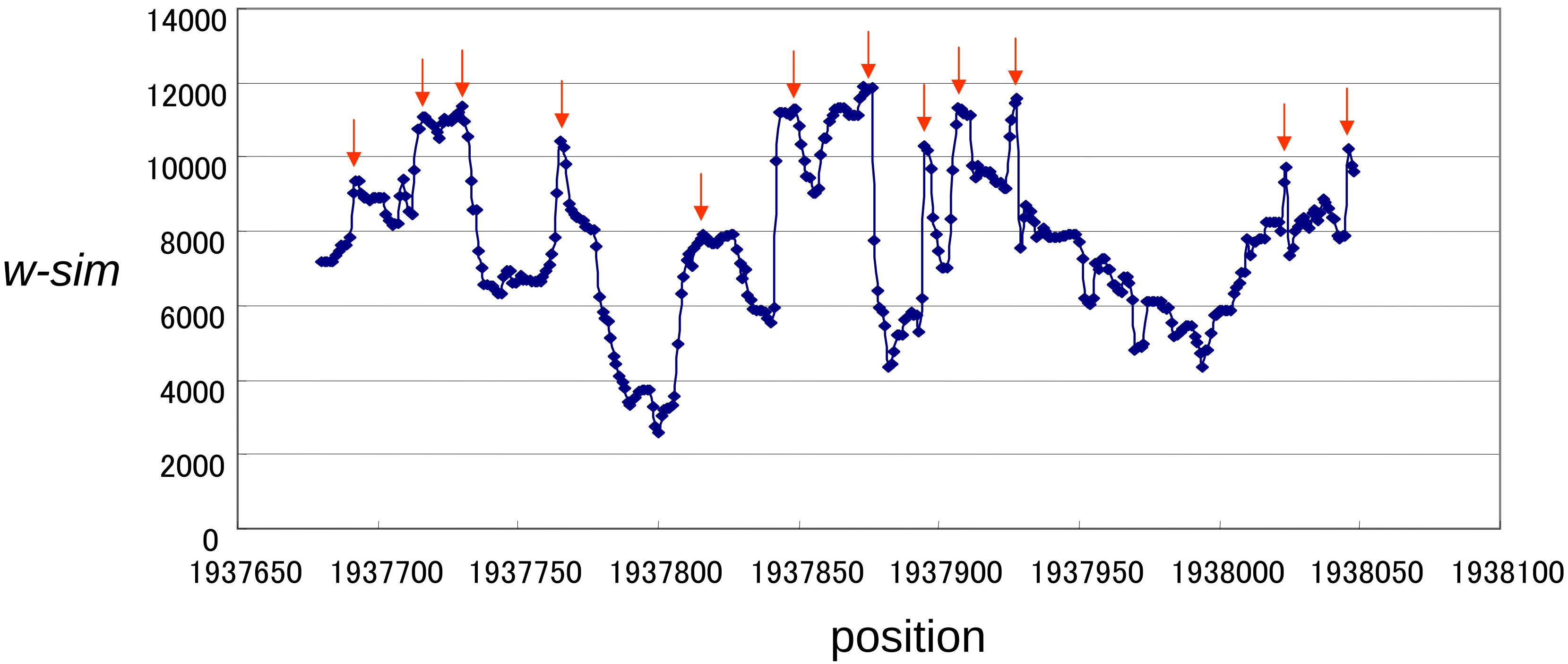
Step 3. Compute subregions R_r on the regulatory region of g_0 such that (i) $max_t\ w\text{-}sim^*(w[g_0, i], g_t)$ is significantly high for almost all $i \in R_r$, and (ii) R_r contains *peak positions* frequently, where $w\text{-}sim^*(w[g, i], g') = max_j\ w\text{-}sim(w[g, i], w[g', j])$.

Step 4. Find a set of transcription factors T_F *dominant* for gene g_0.

Step 5. For each factor $F_k \in T_F$ and each subregion R_r, compute a set of pairs of windows $TFBS(R_r, F_k) = \{ (w[g_0, i], w[g_t, j]) \}$ such that (i) F_k is dominant for gene g_t, (ii) $w\text{-}sim(w[g_0, i], w[g_t, j]) = w\text{-}sim^*(w[g_0, i], g_t)$, and (iii) i is a peak position for g_t.

Peak Positions

- The position i in the regulatory region of gene g is called *a peak position* for gene g' if $i = argmax_{k \in neighbor(i)}\ w\text{-}sim^{*}(w[g, k], g')$.
- It is a local maximum position. The binding site should be contained in windows at peak positions.



Outline of the Procedure

Step 1. Find a set of genes G whose expression patterns are correlated with that of g_0.

Step 2. Compute *window similarity* $w\text{-}sim(w[g_0, i], w[g_t, j])$ for every gene $g_t \in G$ and every positions i, j.

Step 3. Compute subregions R_r on the regulatory region of g_0 such that (i) $max_t\ w\text{-}sim^*(w[g_0, i], g_t)$ is significantly high for almost all $i \in R_r$, and (ii) R_r contains *peak positions* frequently, where $w\text{-}sim^*(w[g, i], g') = max_j\ w\text{-}sim(w[g, i], w[g', j])$.

Step 4. Find a set of transcription factors T_F *dominant* for gene g_0.

Step 5. For each factor $F_k \in T_F$ and each subregion R_r, compute a set of pairs of windows $TFBS(R_r, F_k) = \{ (w[g_0, i], w[g_t, j]) \}$ such that (i) F_k is dominant for gene g_t, (ii) $w\text{-}sim(w[g_0, i], w[g_t, j]) = w\text{-}sim^*(w[g_0, i], g_t)$, and (iii) i is a peak position for g_t.

Dominant Factors

- Find factor F such that the expression level of gene g_0 in gene f-disruption mutant changes significantly, where gene f synthesizes factor F.

araR	4.4235	1
yhjM	3.042298	1
paiB	2.917239	-1
ccpA_V	1.594531	1
acoR	1.402464	-1
sigZ	1.396163	-1
lmrA	1.330305	-1
comK	1.322587	-1
sigF2	1.285625	1
comA	1.240042	-1

log-ratio of araA for each disruption mutant

Experiment

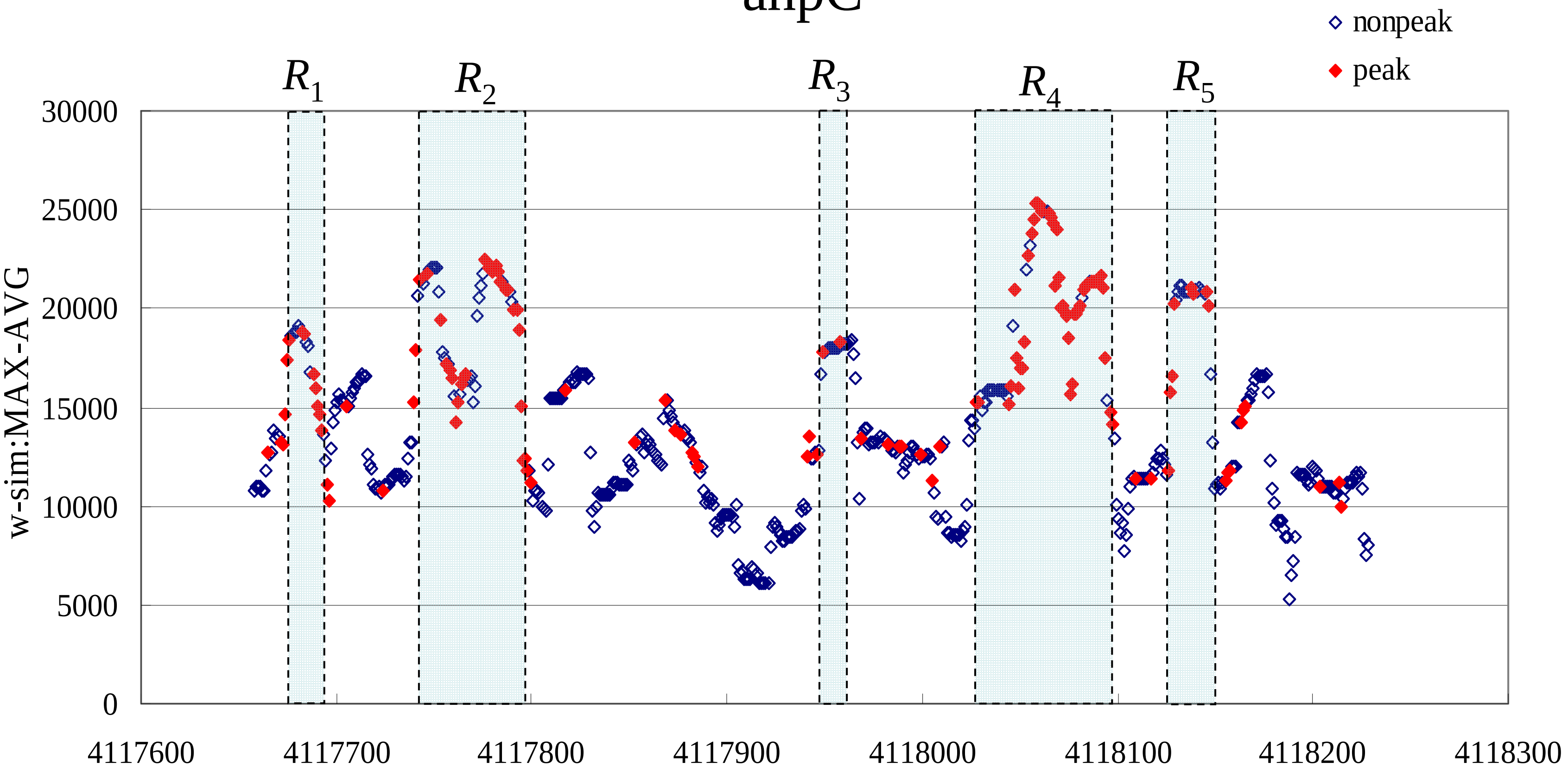
- Genome data: *Bacillus subtilis* (AL009126), and
- Expression data: *Bacillus subtilis*, expression data for 108 gene-disruption mutants.
- g_0 = ahpC.
- We select 400 genes whose expression patterns are correlated with that of ahpC.

Transcription factor: *PerR*

Regulated gene	Operon	Sigma	Regulation	Absolute position	Location	Binding seq.(cis-element)	Exp.	Reference	Year
ahpC	ahpCF	ND	Negative	4118058..4118119	ND	CTTGACAAAAAATATATATTAATTAATAATTCATATATAATTAGAATTATTATTGAAAGCGA	FT	Herbig, A. F., et al.	2001
fur	ND	ND	Negative	2449580..2449594	-49:-35	TTATAATAATTATAG	FT	Fuangthong, M., et al.	2002
hemA	ND	ND	Negative	2878294..2878322	ND	AGAAACTATGTTATAATTATTATAAATAA	FT	Herbig, A. F., et al.	2001
hemA	ND	ND	Negative	2878248..2878289	ND	TTCTATGTTAGAATGATTATAAATTAAGATTGGGTGTTGGGG	FT	Herbig, A. F., et al.	2001
katA	ND	ND	Negative	960520..960577	ND	CTATTTTATAATAATTATAAAATAATATTGACTTTTTACTTAGAGATGATATTATGTT	FT	Herbig, A. F., et al.	2001
mrgA	ND	ND	Negative	3382535..3382589	ND	TCTAAATTATAATTATTATAATTTAGTATTGATTTTTATTTAGTATATGATATAA	FT	Herbig, A. F., et al.	2001
perR	ND	ND	Negative	943933..943958	-13:+13	TTACACTAATTATAAACATTACAATG	FT	Fuangthong, M., et al.	2002
perR	ND	ND	Negative	943942..943964	-4:+18	TTATAAACATTACAATGTAAGAA	FT	Fuangthong, M., et al.	2002
ykvW	ND	ND	Negative	1450655..1450705	-75:-25	TAATGATAATTATTATCAAAAAGAAATTAAAATAATTATAATTGAAATTCT	FT	Gaballa, A., et al.	2002

http://dbtbs.hgc.jp/

ahpC



◇ nonpeak
◆ peak
R_1
R_2
R_3
R_4
R_5
w-sim:MAX-AVG
30000
25000
20000
15000
10000
5000
0
4117600
4117700
4117800
4117900
4118000
4118100
4118200
4118300

TFBS(R_4, PerR)

Factor	g_0	strand	position	g_t	strand	position	*w-sim*
PerR	ahpC	+	4118067	yfmJ	–	818787	41735.37
PerR	ahpC	+	4118069	katA	–	960569	40960
PerR	ahpC	+	4118081	mrgA	+	3382527	35852.64
PerR	ahpC	+	4118081	hemA	–	2878297	34740.15
PerR	ahpC	+	4118068	yfmJ	–	818808	33846.78
PerR	ahpC	+	4118061	yacL	+	108571	31166.68
PerR	ahpC	+	4118081	ykvW	+	1450644	31099.26
PerR	ahpC	+	4118067	glyA	–	3789603	30846.42
PerR	ahpC	+	4118067	yoqS	–	2193550	30812.71
PerR	ahpC	+	4118078	ydjL	–	678938	30644.15
PerR	ahpC	+	4118067	ywdF	–	3897996	30340.74
PerR	ahpC	+	4118067	bmr	+	2493909	30205.89
PerR	ahpC	+	4118066	bmr	+	2493910	29143.97

Estimated Binding Sequences

ahpC: **TAATAATTCATATATAATTAGAATTATTAT**

katA: **TATATCGATTAATAGAGATAACTATTTTAT**

mrgA: **TCAGCTGATCTAAATTATAATTATTATAAT**

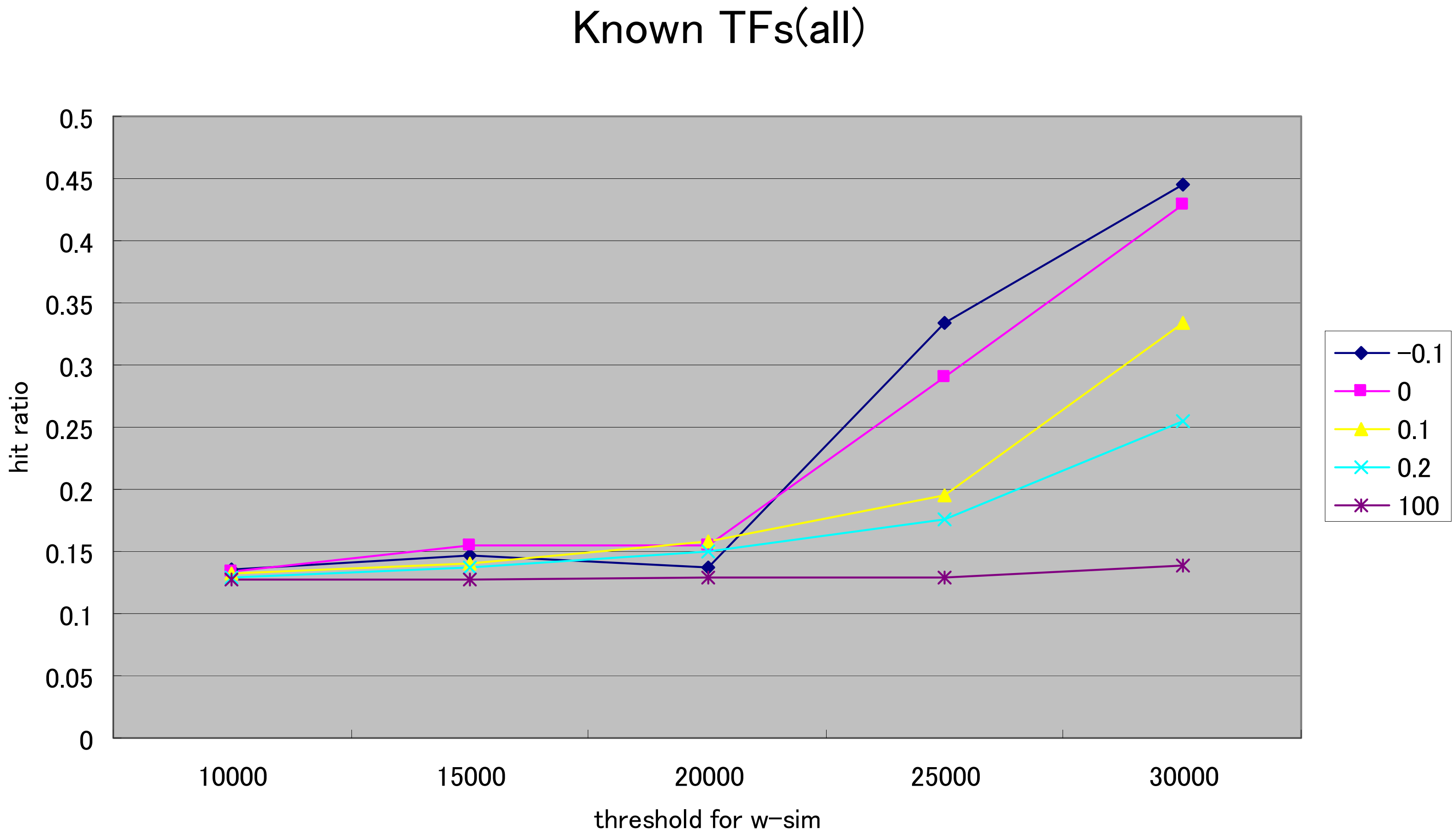
hemA:**TGAAAGAAACTATGTTATAATTATTATAAA**

ykvW: **TGAATAAACATTAATGATAATTATTATCAA**

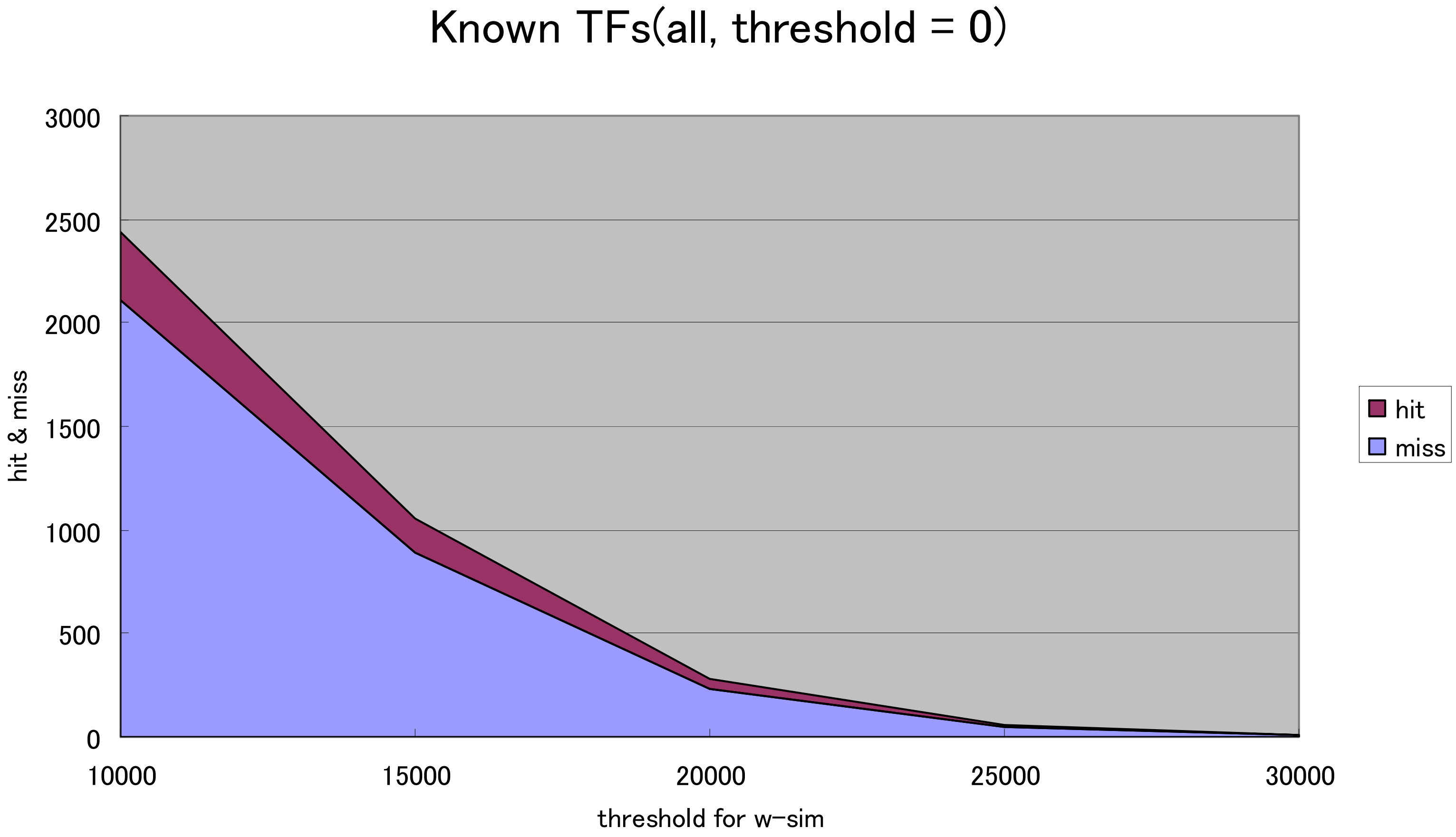
Effect of Sequence Analysis

- For each gene-disruption mutant, we select top 10 genes in the list of genes sorted by decreasing order of log expression ratio.
- If the selected gene is in a operon, we select the first gene in the operon. Then 79 genes are selected.
- We apply the method to the 79 genes and compare the result with known regulation data in http://dbtbs.hgc.jp/.

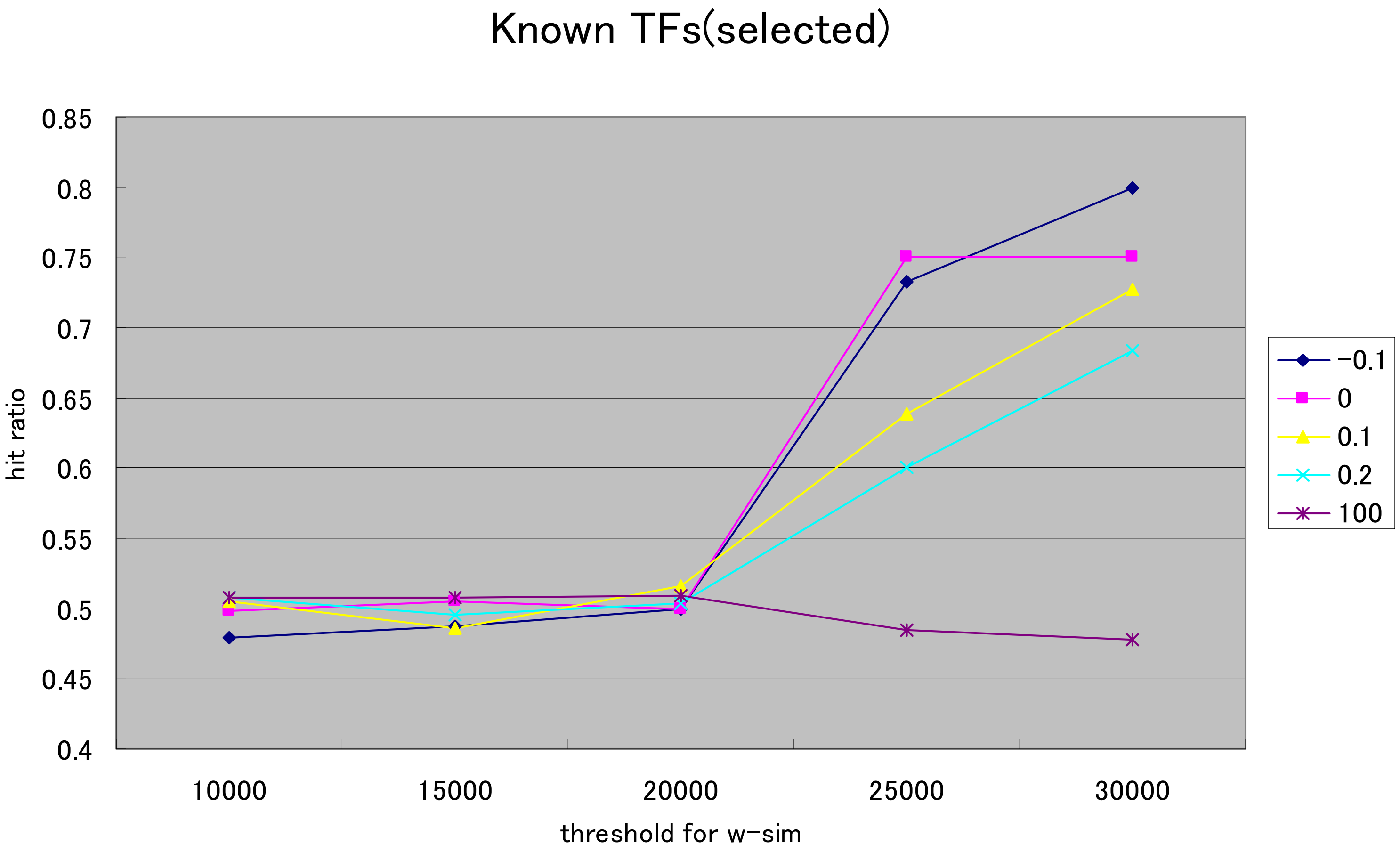
Effect of Sequence Analysis



Effect of Sequence Analysis



Effect of Sequence Analysis



Future Work

- Combination of results using network structure.
- Using databases for known binding sequence.
- Model-based estimation.